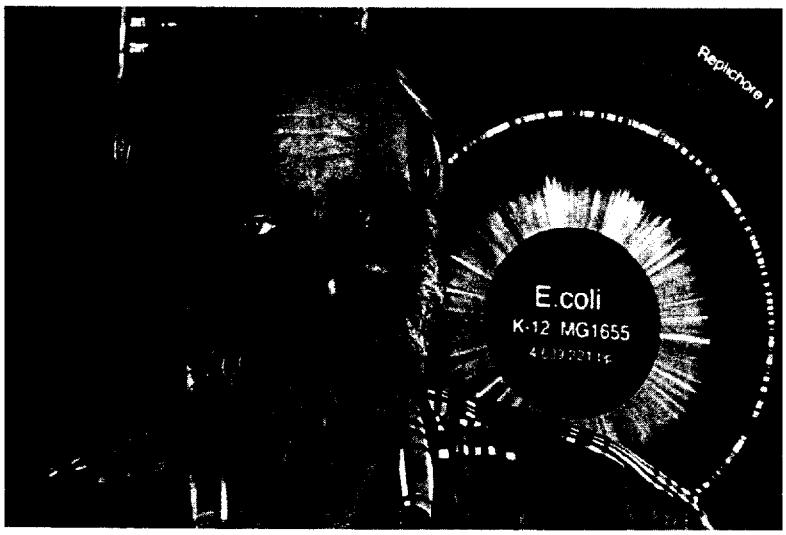
Cracking *E. coli*'s Genomic Code

After years of racing to decode the genetic blueprints for several organisms, scientists are finally reaching the finish line. The latest triumph is the genome, or complete DNA makeup, of the bacterium *Escherichia coli,* unraveled by Dr. Frederick Blattner's team at the University of Wisconsin at Madison.

Genomes of all organisms, ranging from bacteria to humans, are made of only four building blocks called bases that encode the genetic information. The bases are linked together like pearls on a string, but their order varies among the genomes of different organisms. Decoding, or sequencing a genome involves identifying each base along the string. When the base sequences are known, researchers can compare different organisms and obtain information about their biological functions and evolution. Because many genes are similar in primitive and advanced organisms, even bacteria can provide information important for human biomedical research.



Dr. Frederick Blattner and his colleagues have unraveled the complete DNA sequence of the bacterium* Escherichia coli, *which contains about 4,300 genes *(Photo by B. Wolfgang Hoffmann, University of Wisconsin, Madison)*

At a January 1997 meeting about small genomes, Dr. Blattner announced the long-awaited completion of the genomic sequence of the K-12 strain of *E. coli.* The organism contains a total of 4,639,221 base pairs, with about 4,300 genes. The complete annotated sequence is available via the World Wide Web at http://www.genetics.wisc.edu. "The data we have now are of such high quality that there is not one sequence ambiguity for this strain," says Dr. Blattner. "There are likely some errors remaining, but very few."

Begun in 1991, earlier than any other bacterial sequencing endeavor, the *E. coli* genome project proved to be far more time-consuming and complicated than unraveling the genomes of *Hemophilus influenzae, Mycobacterium genitalium, Methanococcus jannaschii,* or *Helicobacter pylori,* which have been published recently. The *E. coli* gene project was like running a marathon with hurdles thrown in for good measure, whereas the others were more akin to 100-yard dashes. For starters, the 4.6 million-base genome of *E. coli* is more than twice as large as each of the genomes of the other four bacteria. Perhaps even more significant, the beginning of Dr. Blattner's genome sequencing project predated the development of automated sequencing techniques and data-management software that were integral parts of the later endeavors. Additional obstacles, in the guise of funding problems, also impeded progress.

The *E. coli* project dates back to 1983, when Dr. Blattner published an article in which he proposed sequencing the entire bacterial genome. "This was even before the inception of the Human Genome Project," he recalls. As the most extensively used vehicle for experiments on bacterial genetics since the 1940s, the K-12 strain of *E. coli* was the obvious choice for such an endeavor.

The sequencing method involved breaking up the genome into smaller pieces, sequencing these fragments at random, and finally, assembling the pieces in correct order with the help of a computer. The first step in this process was to build a library of fragments of the *E. coli* genome, each of which could then be sequenced. By 1991 Dr. Blattner's group had finished this task and was in possession of 450 overlapping pieces, each consisting of 15,000-base fragments of the genome. The next step was to determine the base sequences of all these fragments.

According to Dr. Blattner, "The next years involved massive data gathering by a relatively inexpensive radioactive method." But it was still slow going. At the end of the first three years the researchers had determined the sequence of about 1.6 million bases, approximately a third of the total 4.6 million bases.

Meanwhile, as this work proceeded, other scientists

were developing automated sequencing techniques to speed up the process. Projects to sequence the genomes of other organisms, both large and small, were launched at other institutions, and high-quality sequence data became available at a much faster rate. By that time Dr. Blattner's project was in a catch-22 situation: without renewed funding he could not afford to buy state-of-the-art equipment to upgrade his sequencing effort, but due to the slow progress of his project in comparison to others, funding agencies seemed unwilling to grant him additional awards. Indeed, during part of 1995 his project nearly stopped altogether, while there were speculations by the media that the project might be turned over to other research groups.

The* E. coli *gene project was like running a marathon with hurdles thrown in for good measure. ...

Fortunately, help—in the guise of an NCRR Shared Instrumentation Grant—was around the corner. With this money Dr. Blattner was able to buy automated equipment and proceed with sequencing the genome. "The machines were more rapid and accurate," he explains. "Data management became much simpler because the sequence information was fed directly to computers, which could then look through the reams of data and identify different areas of the DNA chains as genes, promoters, and other functional regions."

Indeed, progress was immediate and dramatic. "In a single year after getting these machines we had completed sequencing the rest of the approximately 3 million base pairs. By mid January, 1997, we submitted a complete annotated version of the sequence to the database," Dr. Blattner says.

Almost two-thirds of the 4,300 genes identified in the K-12 genome have known functions. The others bear no resemblance to known genes or functions. "One of the most exciting things about having the entire sequence is that we can start to discover the functions of unknown genes," says Dr. Blattner. His group is already attempting to compare the genome of *E. coli* K-12, which is nonpathogenic, with a pathogenic *E. coli* strain called 0137:H7 that was isolated from contaminated hamburger meat. The researchers hope to pinpoint genes that are implicated in pathogenicity.

As a resource, Dr. Blattner says, the *E. coli* genome is "more than just data that one can get from the Web site. *E. coli* is the complete sequence of the genome of the organism about which the most is known." From a historical perspective too, it is particularly meaningful that *E. coli* was sequenced at the University of Wisconsin. For many years, the university was the scientific home of Dr. Joshua Lederberg, who won the 1958 Nobel Prize for his pioneering work on bacterial genetics—work largely done using K-12 as well as other strains of *E. coli*. "So in a manner of speaking, *E. coli* was sequenced in its own home," Dr. Blattner says.

The database of *E. coli* gene structure complements another NCRR-supported *E. coli* database called EcoCyc, an encyclopedia of *E. coli* genes and metabolism (see *NCRR Reporter,* May/June 1995, pp. 12–13). EcoCyc, which is a collaborative project of Dr. Peter D. Karp at SRI International in Menlo Park, California, and Dr. Monica Riley at the Marine Biological Laboratories in Woods Hole, Massachusetts, describes *E. coli* metabolic pathways, biochemical reactions, and enzymes. EcoCyc currently describes approximately 130 pathways, more than 700 reactions, and 3,000 genes. EcoCyc is also available on the Internet (see below).

The gene sequencing projects completed to date are themselves major accomplishments, but they are also stepping stones toward the grand prize—unraveling the complete sequence of the human genome. Thanks to the experience and information gained in the smaller races and the continuous development of better equipment, the international Human Genome Project picks up speed as it approaches the finish line.

—Neeraja Sankaran

This work was supported by a Shared Instrumentation Grant from the Biomedical Technology area of the National Center for Research Resources.

More information about E. coli *gene structure is available at http://www.genetics.wisc.edu. The EcoCyc database can be reached at http://www.ai.sri.com/ecocyc/ecocyc.html*

Additional Reading

Blattner, F. R., Plunkett, G., III, Bloch, C. A., et al., The complete sequence of *Escherichia coli* K-12. *Science,* in press.